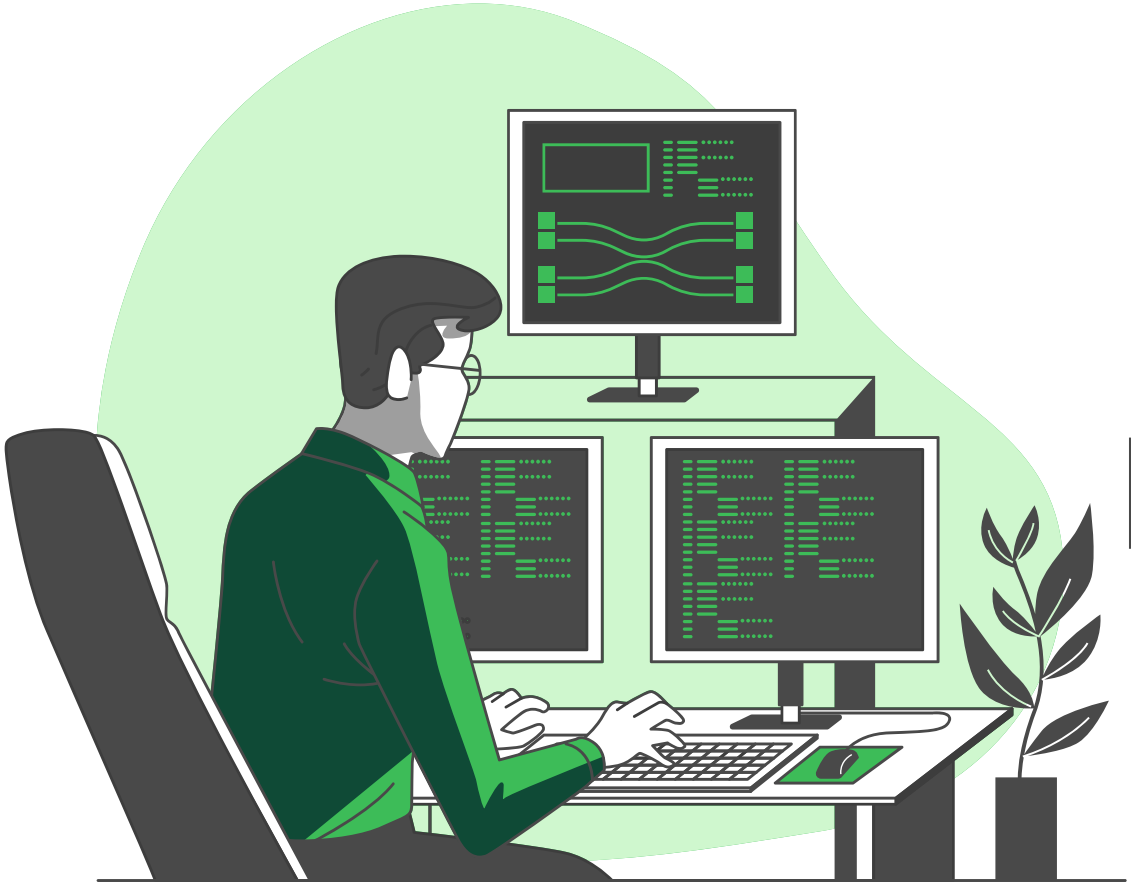
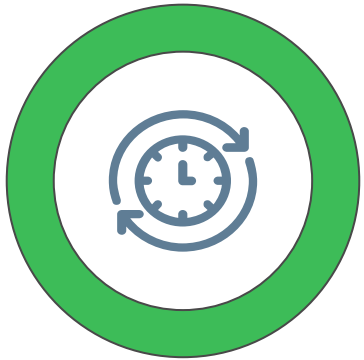


CSC448 Final Project

Team DITTO: Alice Liu, Emily Kim, Lily Liang



CLASSIFICATION PROBLEM

SPAM/NOT SPAM EMAILS

Address the challenge of differentiating between SPAM and NON-SPAM emails

Dataset

Kaggle

Dataset 1
Dataset 2

Dataset Details

Dataset 1:
2 columns, 5k rows
Columns: Category, Message

Dataset 2:
2 columns, 3k rows
Columns: Email, Label

Combining

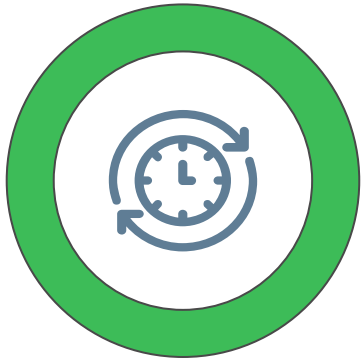
Final Dataset:
2 columns, 8k rows
Columns: Label, Email

Final Dataset

	label	email
0	0	Go until jurong point, crazy.. Available only ...
1	0	Ok lar... Joking wif u oni...
2	1	Free entry in 2 a wkly comp to win FA Cup fina...
3	0	U dun say so early hor... U c already then say...
4	0	Nah I don't think he goes to usf, he lives aro...

0 = non-spam/ham
1 = spam

Email message



WHAT DO YOU WANT TO DO WITH THE DATASET?

Utilizing various machine learning models, we aim to distinguish between spam and non-spam emails, reducing spam impact in inboxes by building an efficient and accurate detection system.

Goals

Accuracy

Achieve a high accuracy rate

Security

Reducing the dangers of phishing emails

Prototype

Create a sample user interface using voila

THANK YOU!

https://github.com/AliceLiu17/csc448_final/tree/main